



# TAEMA : Traitement Automatique de l'Écriture de Mots Affectifs

Pierre-André Buvet, Fabrice Issac

## ► To cite this version:

Pierre-André Buvet, Fabrice Issac. TAEMA : Traitement Automatique de l'Écriture de Mots Affectifs. Verbum ex machina., 2006, 2, pp. 856-867. hal-00680256

**HAL Id: hal-00680256**

**<https://hal.science/hal-00680256>**

Submitted on 20 Mar 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## TAEMA : Traitement Automatique de l'Ecriture de Mots Affectifs

Pierre-André BUVET, Fabrice ISSAC

Laboratoire de Linguistique Informatique CNRS - Université Paris 13, av J.B.  
Clément, 93 430 Villetaneuse

{pierre-andre.buvet, fabrice.issac}@lli.univ-paris13.fr

**Résumé** Nos recherches portent sur l'élaboration et l'utilisation de ressources linguistiques de type dictionnaire. Les dictionnaires utilisés sont capables d'apporter une aide à la rédaction. Ils sont bien plus que des dictionnaires « classiques » puisque l'acte de produire met en oeuvre de nombreuses compétences liant la morphologie, la syntaxe et la sémantique. Nous présentons le prototype d'une application qui génère des phrases simples du français en se basant sur des critères sémantiques. Nous utilisons le modèle des classes d'objet qui est un modèle linguistique qui permet de décrire le lexique du point de vue de ses propriétés syntactico-sémantiques.

**Abstract** Our research is concerned with the development and use of dictionary type resources. Such dictionaries which are of assistance in the composition of texts are considerably more helpful than classical dictionaries. The act of writing requires many integrated skills: syntax, semantic, morphology. We present the prototype of an application enabling to generate simple French sentences based on semantic criteria. We use the model of object classes which is a linguistic model able to manage the lexicon, subjacent syntax and semantics simultaneously.

**Mots-clés :** apprentissage des langues, dictionnaires électroniques, lexique, XML

**Keywords:** language learning, electronic dictionaries, lexicon, XML

## 1 Introduction

Les techniques de TAL sont indispensables à chaque fois que l'objet à traiter informatiquement est de nature linguistique. Il est donc nécessaire d'appliquer ces techniques dans le cadre d'environnements d'apprentissage des langues. L'aide apportée est multiple. Elle peut prendre la forme d'évaluation de productions, de mise à disposition de ressources ou d'incitation à la production via des mises en situations diverses. Nous nous plaçons dans une autre optique. Nous proposons un environnement d'aide à la rédaction basée sur l'utilisation d'une ressource linguistique riche permettant de prendre en compte les descriptions morphologiques, syntaxiques et sémantiques.

Dans un premier temps, nous établirons un parallèle entre l'apprentissage des langues orienté vers le lexique et le traitement des données linguistiques fondé sur des dictionnaires

électroniques pour justifier la spécificité du système d'aide à la rédaction que nous présentons. Dans un deuxième temps, nous présenterons la théorie linguistique qui a été prise en compte pour structurer les dictionnaires électroniques utilisés et nous discuterons des particularités de ces dictionnaires. Dans un troisième temps, une fois décrit le fonctionnement de notre système du point de vue de l'utilisateur, nous ferons état des choix effectués pour rendre compte sur le plan informatique des particularités du modèle.

## 2 Lexique, apprentissage et TAL

Les nombreux travaux existant dans le domaine de l'EAO et plus précisément dans le domaine de l'ALAO montrent que l'ordinateur apporte une aide efficace à l'apprentissage. Cet apport est d'abord d'ordre matériel, lié à la nature même du média informatique, mais aussi cognitif, c'est-à-dire lié à la manière dont l'homme et la machine interagissent.

L'apport principal, le plus visible, concerne les données et les types de données manipulés. En effet, l'ordinateur offre un accès rapide et facile à l'information tant au niveau qualitatif que quantitatif. De plus, le flux d'information est maîtrisé, l'utilisateur peut en effet filtrer l'information suivant des critères qui lui sont personnels. Les interfaces, dont le rôle est de présenter des informations à un utilisateur et de recevoir des ordres de celui-ci, contribuent également à ce que la vigilance de l'apprenant ne soit pas prise en défaut.

Un effort est fait, et cela semble nécessaire dans le cadre de l'apprentissage des langues, pour intégrer des techniques de TAL. Ainsi, les projets Free-Text (L'Haire et al., 2003) et AlexiA (Chanier et al., 1997) proposent des modules évolués d'analyse syntaxique. D'autres projets, tels les plateformes MIRTO (Antoniadis, 2005) et EXILLS<sup>1</sup>, regroupent, outre des ressources, des outils classiques d'analyse textuelles. De ce point de vue, la ressource linguistique proprement dite n'est pas le centre des préoccupations puisque l'objectif est bien souvent de proposer un retour des productions de l'apprenant. D'autres modules sont mis en place : ils permettent l'accès à une ressource riche, l'objectif étant ici d'aider l'apprenant dans sa recherche d'information linguistique. Ainsi AlexiA et le DAFLES (Selva et al., 2003) proposent des dictionnaires dont le contenu et la structure ont été pensés spécifiquement en fonction d'un objectif d'apprentissage.

Nos recherches s'orientent dans cette dernière direction. En effet nous disposons de dictionnaires suffisamment riches pour pouvoir générer des phrases simples (cf. infra). Après avoir présenté l'intérêt d'une orientation lexicale pour l'apprentissage des langues et le TAL, nous indiquons nos vues quant à la mise en œuvre d'un système d'aide à la rédaction utilisant les ressources élaborées à partir du modèle des classes d'objets.

### 2.1 Lexique et apprentissage

L'apprentissage du vocabulaire n'est pas une tâche simple et les informations attachées aux mots sont multiples. Nous allons très rapidement exposer les motivations théoriques pour un apprentissage des langues orienté vers le lexique. Ces motivations sont d'ordre sémantique, syntaxique et pragmatique.

Tout d'abord du point de vue sémantique, on se rend compte que la signification d'un mot n'est pas monovalente, c'est-à-dire que le mot a une propension naturelle à la polysémie. C'est le

---

<sup>1</sup> <http://www.exills.com/>

contexte qui précise les différents sens des mots. Par ailleurs, syntaxe et lexique ne peuvent pas être considérés indépendamment. Même s'il est possible d'utiliser correctement un mot sans utiliser la syntaxe qui lui est propre, cette dernière devient nécessaire dès qu'il s'agit de communiquer à un niveau plus sophistiqué. L'aspect pragmatique en rapport avec le lexique implique la prise en compte de nombreux facteurs qui régissent la sélection du vocabulaire dans un discours : le type de discours, le statut de l'interlocuteur et du locuteur, les motivations, etc.

Un dernier argument en faveur d'un apprentissage orienté vers l'acquisition lexicale est que celle-ci est notoirement insuffisante. Le taux d'acquisition du vocabulaire en langue seconde est assez faible. Ainsi le nombre de mots réutilisables en production est de 1500 après cinq ou six ans d'apprentissage scolaire (Bogaards, 1995).

En résumé l'apprentissage d'un mot fait intervenir de nombreuses connaissances : savoir dans quel contexte on l'utilise à l'oral ou à l'écrit (fréquence d'utilisation, cooccurrences associées), appréhender les limitations de son usage selon les variations de situations et connaître ses comportements morphologiques, syntaxiques et sémantiques.

## 2.2 Lexique et TAL

La polysémie et la polymorphie, d'une part, le figement, d'autre part, sont sources de nombreuses difficultés en TAL. Nous indiquons en quoi ces deux catégories de phénomènes langagiers posent problème.

### 2.2.1 Polysémie et polymorphie

Il est courant de traiter le lexique en distinguant trois niveaux d'analyse : la morphologie, la syntaxe et la sémantique. Un traitement différencié des mots se heurte à deux difficultés : la polysémie et la polymorphie.

La possibilité pour une forme donnée de recevoir des sens différents (<sentiment appréciatif> et <réaction appréciative> en ce qui concerne admiration) est associée à ses particularités syntaxiques (les deux sens de admiration donnent lieu à des complémentations et des constructions différentes : Luc éprouve de l'admiration (pour + \*devant) Luc/Luc éprouve de l'admiration (\*pour + devant) ce tableau). Par conséquent, il vaut mieux recourir à une approche qui tient compte conjointement des propriétés morphologiques, syntaxiques et sémantiques des mots pour faire état de leur caractère polysémique.

La polymorphie est une particularité des seuls prédicats. Un même emploi peut avoir différentes formes. Ainsi, Luc a du mépris pour Léa, Luc méprise Louise et Luc est méprisant envers Léa sont trois phrases équivalentes. Elles sont constituées d'un même prédicat (mépris-) associés aux deux mêmes arguments (Luc et Léa). Chaque forme du prédicat est caractérisée par une construction spécifique. De nouveau, il est clair que seul un traitement conjoint de la morphologie, de la syntaxe et de la sémantique est adéquat pour rendre compte des différentes formes d'un prédicat.

L'importance de la polysémie et celle de sa contrepartie, la polymorphie, ont été sous-estimées en TAL. Ce qui explique de nombreuses insuffisances des systèmes qui opèrent sur des données linguistiques. Une amélioration qualitative des systèmes implique que les variations de sens rattachées à une forme ou les variations de formes rattachées à un sens soient systématiquement traitées. Pour ce faire, nous défendons le point de vue suivant : il faut implémenter dans les systèmes des dictionnaires électroniques à large couverture lexicale qui

permettent de rendre compte de la polysémie et de la polymorphie ; autrement dit, des dictionnaires qui intègrent des descripteurs de nature morphologique, syntaxique et sémantique.

### 2.2.2 Figement

Les expressions -figées correspondent à des séquences de mots dont l'agencement prête peu à la variation et n'est jamais régi par les règles combinatoires usuelles. Il s'ensuit que leur interprétation n'est généralement pas déductible de celle de leurs constituants. Ainsi, le fait que la suite être -bouche bée devant est un synonyme de admirer n'est pas directement explicable à partir du sens des mots qui la constituent. Par ailleurs, de nombreuses expressions figées peuvent correspondre à une séquence libre (typiquement prendre le taureau par les cornes) de telle sorte qu'il est difficile de déterminer quelle interprétation privilégiée sur la seule base de l'assemblage de leurs formes.

Nous considérons que la prise en compte des expressions figées en TAL -implique qu'elles fassent l'objet de recensements exhaustifs et systématiques. La seule notion de collocation (au sens de grande fréquence de proximité entre au moins deux mots) -est insuffisante car elle rassemble des données très hétérogènes. Il faut notamment distinguer les expressions figées selon qu'elles procèdent de la signification lexicale ou de la signification grammaticale (Blanco et Buvet 2004).

Dans le premier cas, elles fonctionnent comme des adjectifs, des noms ou des verbes ; il est impératif alors de les répertorier dans des bases de données et d'y indiquer leurs particularités morphologiques, leurs propriétés distributionnelles et les classes sémantiques qui les caractérisent. Par exemple, la séquence prendre la poudre d'escampette est catégorisée comme une -locution verbale dont le domaine d'arguments unaire se rapporte nécessairement à un humain et est rattachée à la classe sémantique <déplacement humain>.

Dans le second cas, il s'agit principalement de séquences déterminatives (comme un éclair de dans Luc a eu un éclair de génie) ou adverbiales (par exemple comme un pinson dans Luc est gai comme un pinson). Leur signification peut être rapportée à un petit nombre de valeurs ('quantité forte', 'quantité faible', 'intensité forte', 'intensité faible', 'mélioratif', 'péjoratif', 'fréquentatif', etc.). Généralement, la portée distributionnelle de ces séquences est réduite puisqu'elles se combinent avec un nombre très limité de noms, d'adjectifs ou de verbes (\*Luc a eu un éclair de esprit ; \*Luc est content comme un pinson)-. Pour autant, leur association avec d'autres mots ne correspond pas à des expressions figées puisque ces mots peuvent apparaître sans les séquences en question (Luc a (un éclair de + du) génie ; Luc est gai (E + comme un pinson)). Les séquences déterminatives et adverbiales figées doivent être également toutes recensées et il convient d'expliquer les particularités de leur mode de fonctionnement (notamment avec quels items elles sont compatibles et quelles sont leurs valeurs).

## 2.3 Mise en place d'un SAR

L'importance de la part du vocabulaire dans l'apprentissage d'une langue et la possibilité en TAL -d'améliorer les systèmes en implémentant des dictionnaires électroniques à large couverture constitue un point de convergence. Cela nous a conduits à développer un système d'aide à l'apprentissage qui met en avant la dimension lexicale des langues. D'autant plus que le FLE (Français Langue Etrangère) et le TAL ont des problématiques communes vis-à-vis des mots : le traitement de la polysémie, de la polymorphie et du figement sont les mêmes pour un apprenant -ou un système opérant sur des données linguistiques. La nécessité de

prendre en compte conjointement les propriétés morphologiques, syntaxiques et sémantiques des mots pour expliquer leur mode de fonctionnement est valable dans l'un et l'autre cas.

Le système d'aide à la rédaction développé s'intitule TAEMA (Traitement Automatique de l'Écriture de Mots Affectifs). Il s'agit de produire des phrases centrées sur le vocabulaire affectif du français pour des apprenants en langue seconde ou pour des apprenants natifs. Un affect est défini comme un état psychologique (par exemple, la 'joie', la 'peur', 'l'amour', le 'regret') qui est ressenti et qui n'est pas inhérent à un individu (cf. infra). La particularité du dictionnaire des prédicats d'<affect> utilisé permet au système non seulement d'indiquer tout le vocabulaire en rapport avec un type d'affect donné (par exemple une émotion ou un sentiment particulier) mais aussi toutes les constructions associées avec ce vocabulaire. Au final, après avoir fait une requête relative à un type d'affect, l'utilisateur a la possibilité de choisir parmi un ensemble de phrases équivalentes celle qui lui semble la plus adéquate pour sa production écrite.

Avant de présenter TAEMA, nous discutons des particularités linguistiques des ressources lexicales utilisées dans le système.

### 3 Les données linguistiques : théorie et applications

Le modèle des classes d'objets a donné lieu à la réalisation de dictionnaires électroniques destinés aux divers systèmes opérant sur des données linguistiques. Ces dictionnaires visent à une couverture exhaustive du français, entre autres langues. Il s'agit de décrire le lexique avec des propriétés syntactico-sémantiques explicites et reproductibles qui sont susceptibles de faire l'objet de procédures informatisées.

Nous présentons en premier lieu le modèle des classes d'objets. Nous illustrons ensuite la description du lexique qui en résulte avec les prédicats d'<affect> dans la mesure où ce sont ces données lexicales qui sont implémentées dans le système TAEMA. Nous terminons par une présentation du mode de structuration des dictionnaires électroniques.

#### 3.1 Le modèle des classes d'objets

Le modèle des classes d'objets (Gross 1995, Gross 1996, Le Pesant et Mathieu-Colas 1998) postule que toute phrase est constituée d'un prédicat et de ses arguments et que les autres unités linguistiques ressortissent à l'actualisation. Les prédicats prédominent structurellement les arguments et les conditions d'apparition des différents actualisateurs sont subordonnées aux différentes relations entre les prédicats et les arguments constitutives de phrases.

Le modèle a comme conséquence que la partition des unités linguistiques en fonction de leur statut de prédicat, d'argument élémentaire ou d'actualisateur ne recoupe pas celle qui a trait aux parties du discours traditionnelles. Ainsi, les prédicats peuvent correspondre, entre autres, à des verbes (chérir dans Luc chérit Léa) ; des adjectifs (épris dans Luc est épris de Léa) ; des noms (béguin dans Luc a le béguin pour Léa). De même, les verbes peuvent être soit précatifs (gifler dans Léa a giflé Luc) soit supports (donner dans Léa a donné une gifle à Luc).

Le modèle des classes d'objets permet de prendre en compte la polysémie. Ainsi, pour ce qui est des prédicats verbaux, il est fait état de leurs schémas d'arguments. Cela a conduit en particulier à factoriser les arguments en les typant sémantiquement en termes de classes. Pour chaque emploi identifié sans ambiguïté, la description gagne en pertinence à tous les niveaux



d'analyse : la conjugaison (certains emplois se distinguent par des particularités de temps, de nombre ou de personne), la dérivation (le lien entre les formes est dépendant du sens), la synonymie et les restructurations (passif, extraction, etc.). Par exemple, la forme conduire a autant d'emplois que de compléments nominaux définis soit en intension (par le biais de traits ou de classes d'objets) soit en extension : conduire un ami à l'hôpital (conduire1 = emmener) ; conduire le troupeau à l'abreuvoir (conduire2 = mener) ; conduire l'autocar (conduire3 = piloter) ; conduire la forêt (conduire4 = aménager) ; conduire la chaleur (conduire5 = véhiculer) ; etc.

Les classes d'objets constituent des ensembles d'items sémantiquement homogènes définis à l'aide de propriétés syntaxiques. On distingue les classes d'arguments, d'une part, les classes de prédicats, d'autre part. Les premières résultent d'une sous-catégorisation syntactico-sémantique des substantifs correspondant aux arguments élémentaires. Les secondes ont trait essentiellement à des adjectifs, des noms et des verbes. Ce classement est pris en compte dans la description des prédicats en termes de classes. Il s'agit cependant d'une catégorisation effectuée sur la base de leurs propriétés syntaxiques et sémantiques remarquables. De ce fait, la caractérisation syntactico-sémantique d'un prédicat prime sur ses particularités morphologiques lorsqu'il recouvre deux de ces formes, voire les trois. Ainsi, les phrases Luc a désiré cela, Luc a été désireux de cela et -Léa a eu le désir de cela sont considérées comme strictement équivalentes dans la mesure où les arguments Luc et cela se rapportent à un même prédicat qui recouvre soit la forme verbale (désirer) soit la forme adjectivale (désireux) soit la forme nominale (désir). -Le modèle rend donc également compte de la polymorphie.

Depuis de nombreuses années, les séquences figées du français ont fait l'objet de recensements exhaustifs. Plus de 100 000 noms composés de la langue générale ont ainsi été recueillis. Parallèlement, un travail similaire est en cours pour les langues de spécialité. D'autres travaux portent sur les adjectifs et les verbes composés (à ce jour 10 000 locutions adjectivales et 30 000 locutions verbales ont été listées). Les recensements exhaustifs et systématiques entrepris recouvrent une grande partie des constructions figées du français. Leur spécification dans les dictionnaires électroniques est une contribution majeure au traitement du figement. Ainsi, le système INTEX (Silberztein 1993) intègre -l'ensemble de ces dictionnaires.

### 3.2 Les prédicats d'<affect>

Nous présentons rapidement les prédicats d'<affect> en termes de classes d'objets dans la mesure où c'est le vocabulaire qui a été implémentée dans le système TAEMA.

La langue française dispose de toutes sortes d'adjectifs, de noms et de verbes pour exprimer ce qui se rapporte à l'intériorité mentale des êtres humains, une grande partie de ce lexique étant considérée par les linguistes comme des termes psychologiques.

L'intériorité mentale se scinde en deux selon sa nature cognitive ou bien psychologique. -Il est possible de répartir les prédicats de la seconde catégorie comme suit : (i) les prédicats de <disposition d'esprit> (colérique) ; (ii) les prédicats d'<appréciatif psychologique> (désobligeant) ; (iii) les prédicats d'<affect> (amoureux).

Une justification rapide de cette sous-catégorisation tient en deux points, plus précisément-deux oppositions. La première opposition se manifeste comme suit : les prédicats de la

première sous-catégorie correspondent à des états psychologiques inhérents, -les prédicats des deux autres sous-catégories à des états psychologiques contingents. Le caractère permanent ou occasionnel de l'état auquel se rapporte un prédicat explique l'incompatibilité ou la compatibilité du prédicat avec des adverbes à valeur durative. Ainsi, constamment se combine difficilement avec le prédicat dans le premier cas, la caractérisation aspectuelle imputable à l'adverbe est superfétatoire par rapport au trait permanent du prédicat, mais beaucoup plus facilement dans le second cas, l'adverbe annihile le trait occasionnel du prédicat en indiquant un aspect duratif : ?\_Luc est constamment— peureux : prédicat du type (i) ; Luc est constamment sévère avec Max : prédicat du type (ii) ; Luc est constamment apeuré : prédicat du type (iii).

La deuxième opposition résulte du fait que les prédicats de la troisième sous-catégorie ont comme particularité d'être en rapport avec un 'ressenti intérieur' alors que les prédicats des deux autres sous-catégories ont trait à un 'jugement extérieur'. Les verbes d'<opinion>, e.g. trouver, rendent compte de la distinction en admettant plus facilement les complétives formées à partir d'un prédicat de <disposition d'esprit> ou de <comportement> que celles formées à partir d'un prédicat d'<affect> : Je trouve que Luc est (avenant + acariâtre) : prédicats du type (i) ; Je trouve que Luc est (bienveillant + désobligeant) avec Max : prédicats du type (ii) ; ?Je trouve que Luc est (amoureux de Léa+ dégoûté de la vie) : \_prédicats du type (iii).

En résumé, les prédicats d'<affect> sont des adjectifs, des noms et des verbes en rapport avec des états psychologiques contingents qui sont de l'ordre du ressenti. Le champ d'étude lexicale une fois précisé, une nomenclature a été établie.

### 3.3 Typologie et structuration des dictionnaires du LLI

L'exploitation des dictionnaires traditionnels (typiquement Le Robert) en TAL pose de nombreuses difficultés étant donné qu'ils ne sont pas complets (seuls les sens les plus fréquents sont traités), que les descriptions linguistiques sont insuffisantes (toutes les informations nécessaires à la construction des mots ne sont pas spécifiées et celles relatives à l'aspect inhérent ou contextuel sont totalement ignorées), que les classifications sémantiques sont peu cohérentes (on peut difficilement retrouver les mots qui appartiennent à la même classe). Il est plus aisé d'implémenter dans les systèmes qui opèrent sur des données linguistiques des dictionnaires électroniques correspondant à des bases de données où une entrée lexicale est associée à des descripteurs normalisés (Courtois et Silberztein 1990).

Les dictionnaires électroniques pris en compte ici ont les particularités suivantes : -(i) ils rapportent les unités lexicales à la distinction prédicats de premier ordre/arguments élémentaires (cela a pour conséquence que les différentes formes d'un même prédicat sont décrites de la même façon) ; (ii) les descripteurs associés à chaque entrée ne sont pas hiérarchisés (ils sont de nature morphologique, syntaxique et sémantique) ; (iii) les descripteurs se rapportent au modèle des classes d'objets.

Le recours à des dictionnaires électroniques dans un système d'aide à la rédaction permet de produire toutes sortes de phrases afin d'aider les rédacteurs à -améliorer leur production.

## 4 Système d'aide à la rédaction de mots affectifs

L'aide à l'écriture, d'un point de vue informatique, peut être envisagée à partir de plusieurs fonctions :



- la planification : recherche d'idées, dictionnaires, banques de textes, générateur de textes ;
- l'édition : dactylographie, couper/copier/coller ;
- la correction : recherche/remplacement, correcteur orthographique et grammatical ;
- la présentation : mise en page, maquetage, impression ;
- la collaboration : écriture en chaîne, écriture collaborative, discussion et document partagé.

Par ailleurs, la notion d'interactivité est primordiale (Mangenot 2000). Elle fait partie des critères d'évaluation des outils et peut être plus ou moins développée. Dans cette optique, on peut proposer des activités d'écriture basées sur le dialogue (Caviglia et al., 1994). Cependant le dialogue repose sur les types du discours (narration, argumentation, ...) et non pas, comme nous le proposons, sur un modèle linguistique permettant de rendre compte à la fois de la morphologie, de la syntaxe et de la sémantique.

Nous allons maintenant présenter l'interface du système et la structure interne utilisée pour rendre compte de la richesse du modèle linguistique.

#### 4.1 L'interface de TAEMA

Le prototype développé est susceptible de s'adresser aussi bien à des apprenants L1 que des apprenants L2 pour peu que ces derniers aient un niveau de langue minimum. En effet, l'utilisation de l'interface nécessite une connaissance sommaire du français. L'usage du logiciel présuppose que les utilisateurs sachent exprimer au moins la phrase prototypique associée à la notion et l'environnement spécifié (par exemple Luc aime Léa pour l'affect <amour>). La finalité du logiciel est de lui proposer toutes les paraphrases possibles (Luc a le béguin pour Léa, Luc en pince pour Léa, Luc est fou de Léa, etc.) Il est cependant à noter que nous présentons un prototype non scénarisé et qu'il conviendrait de l'adapter à l'apprenant.

Du point de vue de l'utilisateur du système, il s'agit d'indiquer un type d'affect donné et des éléments contextuels afin d'obtenir toutes les phrases canoniques du français qui sont en rapport avec les indications fournies. Le système développé permet –à un utilisateur de sélectionner (sous forme de menus déroulants ou en les spécifiant en langue naturelle) les différents éléments d'une phrase à produire. ~~Nous prenons en compte deux types d'utilisateurs : des apprenants avancés en français langue seconde ou des jeunes apprenants (niveau cours moyen) natifs.~~ Nous détaillons les différents écrans qui permettent à l'utilisateur de formuler sa requête.

Le premier écran présente : (i) la finalité du projet ; (ii) un mode d'emploi ; (iii) un bouton 'entrée' qui permet d'accéder au système.

Le deuxième écran invite l'utilisateur à choisir un type d'affect dans un champ de saisie à l'aide d'un menu déroulant. Il s'agit de l'appellation de l'une des sous-classes de prédicats d'<affect> qui figurent dans le dictionnaire électronique utilisé par le système. Cette page comporte également trois autres champs de saisie qui ne sont pas activés à ce niveau d'utilisation. Le premier est dit 'expérienceur' : il s'agit de la personne qui ressent l'affect. Le second est dit 'bénéficiaire/objet' : il s'agit de l'éventuelle personne ou entité sur qui porte l'affect. Le troisième est dit 'temps' : il permet de conjuguer le prédicat à un temps grammatical simple.

Le troisième écran rend actifs en partie ou en totalité les champs de saisie autres que celui déjà rempli. Selon que le prédicat est monadique ou dyadique (s'il comporte un ou deux

arguments), les seuls champs 'expérienceur' et 'temps' ou ces deux champs ainsi que le champ 'bénéficiaire/objet' sont disponibles. Les champs 'expérienceur' et 'bénéficiaire/objet' présentent deux options : d'une part, si l'utilisateur souhaite que les arguments correspondent à des pronoms, il fait appel à un menu déroulant où il spécifie la personne, le nombre, le genre du pronom désiré ; d'autre part, -s'il veut que les arguments soient des expressions (groupes nominaux, complétives, infinitives ou entités nommées), il est invité à remplir lui même le champ tout en spécifiant le nombre et le genre de l'expression.

Le quatrième -écran donne tous les résultats -relatifs à la requête de l'utilisateur de telle sorte qu'il puisse choisir une des phrases relatives à l'affect spécifié et les éléments contextuels qu'il a précisé. Le prototype basé sur environ 60 sous-classes permet de générer à peu près 3000 phrases simples du français.

## 4.2 Nécessité de structuration des données

Les ressources élaborées sont structurées via XML. Ce choix permet d'utiliser une gamme très large d'outils mais surtout de définir une ressource très riche qui peut servir de format pivot. Cette ressource pouvant être déclinée en fonction du support ou de l'objectif à atteindre.

Il faut noter que nous utiliserons plusieurs DTD plutôt qu'une seule pour affiner au maximum le balisage. Jusqu'à présent, nous avons établi que le dictionnaire des prédicats nécessitait au moins cinq DTD (chacun repéré par un espace de nom spécifique).

1. En-tête  $\_$  : -contient les informations portant sur l'utilisateur ayant fait la dernière modification, mais aussi la date de la dernière modification, le nombre de révisions... On s'inspirera pour cet en-tête du travail réalisé pour la XCES (XML Corpus Encoding Standard). Cette partie permet d'envisager un développement coopératif d'une part et de gérer l'historique de l'évolution des dictionnaires ;
2. Le document maître  $\_$  : son rôle est de regrouper les différentes parties du dictionnaire ;
3. Les prédicats ;
4. Les arguments ;
5. Hiérarchie  $\_$  : décrit le lien qui existe entre les classes et les hyperclasses.

La figure 1 décrit la manière dont les différentes informations se combinent. Deux DTD supplémentaires sont envisagées pour l'instant et concerneraient la structuration des déterminants et des prépositions.

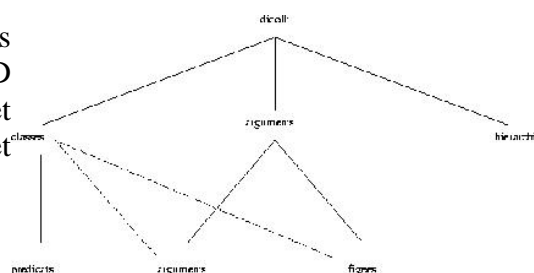


Figure 1 : organisation des différents composants des dictionnaires

Le nombre de DTD peut sembler important, mais ce découpage permet de résoudre un certain nombre de problèmes. En effet, un argument pouvant appartenir à différentes classes d'argument de par la polysémie, il est nécessaire que cette information soit déportée pour éviter un phénomène de redondance. Pour les mêmes raisons, l'hyperclasse d'une classe d'argument, qui selon le schéma d'argument du prédicat peut varier, doit être déportée. Ces possibilités doivent donc apparaître au sein de la structure du dictionnaire. Un autre argument en faveur du découpage concerne la mutualisation des informations par les différents concepteurs des dictionnaires. En effet, la constitution des dictionnaires de prédicats nécessite

de connaître l'ensemble des classes d'arguments afin de rédiger le schéma ; de même la hiérarchie des classes doit être mutualisée.

Exemple de description du nom commun *artère* dans le document XML décrivant les arguments :

```
<elt type = "arg">
  <nom>artère</nom>
  <liste_classe>
    <classe id="voie"/>
    <classe id="partie interne corps"/>
  </liste_classe>
</elt>
```

Le substantif *artère* est un nom commun polysémique. Ce nom a deux sens et appartient donc à deux classes d'argument : <les voies de communication>, notée ici <voie> et les <éléments internes du corps d'un animal ou d'un humain>, notée ici <partie interne corps>.

Exemple de description du nom commun *visage* :

```
<elt type = "arg">
  <nom>visage</nom>
  <liste_classe>
    <classe id="npc"/>
  </liste_classe>
</elt>
```

Exemple de description du schéma d'argument du prédicat verbal *porter* :

```
<schema>
  <predicat>porter</predicat>
  <comp type = "sujet">
    <hyperclasse id="humain"/>
  </comp>
  <comp type = "COD">
    <hyperclasse id = "inc"/>
    <classe id = "vetement"/>
  </comp>
</schema>

<schema>
  <predicat>porter</predicat>
  <comp type = "sujet" pos="0">
    <hyperclasse id="humain"/>
  </comp>
  <comp type = "COD" pos="1">
    <hyperclasse id = "ina"/>
    <classe id = "appellation"/>
  </comp>
</schema>
```

Le prédicat verbal *porter* est polysémique. Chaque sens possède un schéma d'arguments spécifique. Ainsi, le premier schéma d'argument présenté ici -correspond à une phrase du type *Jeanne porte une jupe* tandis que la seconde- phrase correspond à une phrase du type *Emeric porte un nom breton*.

## 5 Conclusions/Perspectives

Nous avons présenté le prototype d'une application permettant de générer des phrases simples du français. L'objectif étant de fournir à un apprenant un outil lui permettant d'appréhender le contexte associé au lexique. L'organisation des informations dans le cadre de ce prototype est primordiale. Nous avons utilisé le modèle des classes d'objet qui est un modèle linguistique

permettant de rendre compte, comme nous le souhaitons du lexique du point de vue de ses propriétés syntactico-sémantiques.

Notre objectif initial était de valider le modèle linguistique, tant du point de vue théorique que fonctionnel, dans le cadre d'un outil d'aide à la rédaction. Il nous a donc fallu définir les limites de ce qui pouvait être utilisé, sans que le modèle en soit affecté. Ce travail a été réalisé de paire avec l'élaboration d'un modèle informatique qui permet de représenter toute la richesse de l'information linguistique. Suite à ce travail théorique, il était nécessaire de valider les ressources proprement dites. Les copies d'écran ci-dessous montrent un exemple d'utilisation du prototype<sup>2</sup>.

### TAEMA

#### Choix de l'affect

Classe :

- admiration
- antipathie\_sympathie**
- colère
- crainte
- dégout
- désespérance
- émotion
- envie
- fierté
- honte
- indignation
- irritation
- mauvaise humeur
- mépris
- pitié
- plaisir
- respect
- satisfaction
- stupefaction

### RESULTAT

***Vous avez choisi l'affect: antipathie\_sympathie***

[retour au flexionneur](#)

**Prédicats nominaux :**  
antipathie  
sympathie

**Phrases de Taema**  
Luc témoigner de la sympathie pour Marie  
Luc montrer de la sympathie pour Marie  
Luc manifester de la sympathie pour Marie  
Luc exprimer sa sympathie pour Marie  
Luc être débordant de sympathie pour Marie  
Luc avoir de la sympathie pour Marie  
Luc éprouver de la sympathie pour Marie  
Luc ressentir de la sympathie pour Marie  
Luc témoigner de l'antipathie pour Marie  
Luc nourir une antipathie pour Marie  
Luc avoir de l'antipathie pour Marie  
Luc éprouver de l'antipathie pour Marie  
Luc ressentir de l'antipathie pour Marie

La réalisation de ces trois aspects du projet s'est fait en parallèle et de manière itérative puisque toute incohérence ou difficulté nous obligeaient à adapter les modèles linguistiques et informatiques sous-jacents. Bien évidemment, le prototype réalisé reste très incomplet en ce qui concerne sa couverture (il permet néanmoins de générer l'ensemble des phrases en rapport avec le lexique de l'affect).

Le travail que nous devons maintenant entreprendre concerne tout à la fois la couverture linguistique et l'interface avec l'utilisateur. Nous avons en effet focalisé nos efforts sur la pertinence des ressources. La présentation de ces ressources, qui sont par nature très complexes, reste un point de réflexion important. Couplé à ce module d'aide à la rédaction, il nous faut envisager à présent une scénarisation pédagogique propre à guider les apprenants.

=

<sup>2</sup> Le prototype présenté ici n'intègre pas encore le module de flexion morphologique. La version finale permettra d'obtenir par exemple « Luc témoigne de la sympathie pour Marie ».

## Références

~~Références (Antoniadis 2005), références EXILLS2~~ ANTONIADIS G., ECHINARD S., KRAIF O., LEBARBÉ T., PONTON C. (2005). « Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO ». *ALSIC* vol. 8, pp. 65-79.

BLANCO X., BUVET P.-A. (2004). « Verbes supports et significations grammaticales. Implications pour la traduction espagnol-français ». *Lingvisticae Investigationes* 27 (2) , pp. 327-342.

BOOGAARDS P. (1995). *Le vocabulaire dans l'apprentissage des langues étrangères*. LAL, CREDIF, HATIER/DIDIER.

BUVET P.-A., GIRARDIN C., GROSS G., GROUD C. (2005). « Les prédicats d'<affect> ». *LIDIL* 32 , pp. 125-143.

CAVAGLIA F., FERRARIS M. (1994). « Scrivere in collaborazione con il computer », dans Mangenot, Camilleri, Poletti, ed. p. 113-119.

CHANIER T., FOUQUERÉ C., ISSAC F. (1997). « AlexiA : un environnement d'aide à l'apprentissage lexical du français langue seconde ». *La locution : entre lexique syntaxe et pragmatique*, Fiala P., Lafon P., Piguet M.-F.

COURTOIS B., SILBERZTEIN M. (1990). « Dictionnaires électroniques du français », *Langue française* 87, Paris : Larousse.

~~GROSS G. (1995). « Une sémantique nouvelle pour la traduction automatique : les classes d'objets », *La Tribune des Industries de la Langue et l'Information électronique*, 17-18-19, Paris, pp. 16-19.~~

GROSS G. (1996a). « Les expressions figées en français : noms composés et autres locutions, Ophrys », Paris-Gap.

GROSS G. (1996b). « Prédicats nominaux et compatibilité aspectuelle » *Langages*, 121, Larousse, Paris, pp. 54-72.

L'HAIRE S., VANDERVENTER FALTIN A. (2003). « Diagnostique d'erreurs dans le projet Free-Text », *ALSIC*, vol. 6, n°2.

LE PESANT D., MATHIEU-COLAS M. (1998). « Introduction aux classes d'objets », *Langages*, 131, Larousse, Paris, pp. 6-33.

MANGENOT F. (2000). « Aide à l'écriture ou environnements d'écriture ? », *Lecture-Ecriture et Nouvelles Technologies*, Coord. J. Anis et N. Marty, p59-68, La collection de l'ingénierie éducative, CNDP, Paris.

MEJRI S. (1997). « Le figement lexical : descriptions linguistiques et structuration sémantique », *série linguistique X*, Publications de la Faculté des lettres de la Manouba.

SELVA T., VERLINDE S., BINON J. (2003). « Vers une deuxième génération de dictionnaires électroniques, Traitement Automatique des Langues ». *Les dictionnaires électroniques*, vol. 44, n° 2, pp 177-197.

SILBERZTEIN M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Masson, Paris.